Estimation bayésienne des matrices de corrélation: Stratégies de formulation de lois a priori

Rapport de recherche présenté

par

Anderson Walter Nzabandora

au

Département de sciences économiques

en vue de l'obtention

du diplôme de

Maîtrise en Sciences (M.Sc.)

en

Sciences économiques, Option: Économie financière

Université de Montréal

Montréal, Québec

Janvier, 2008

Résumé

Ce rapport de recherche développe une méthodologie d'estimation bayésienne des matrices de corrélation. Notre modélisation repose sur la possibilité d'une factorisation symétrique des matrices définie-positives, permettant ainsi de formuler des lois a priori sur le "facteur" de la décomposition, qui tient lieu de nouveau paramètre du modèle.

La formulation de croyances a priori sur le degré de corrélation des différentes variables se fait au moyen d'un ensemble d'hyperparamètres angulaires. De ce fait, notre modélisation exploite largement l'interprétation géométrique du concept de corrélation, en termes de produit scalaire usuel des vecteurs-lignes du nouveau paramètre factoriel.

Différentes variantes du modèle sont proposées afin de tenir compte des trois principaux cas de figure de Liechty (2004), auquel le présent travail se veut une alternative. Il s'agit: d'un ensemble de variables présentant, par paires, des degrés de corrélation identiquement distribués; d'un ensemble de variables affichant des corrélations différentes entre elles; et en dernier lieu, d'un ensemble de variables subdivisé en groupes selon l'homogénéité interne et l'hétérogénéité externe des différents groupes, en termes de corrélation.

En focalisant notre étude sur une famille particulière de distributions a priori, nous proposons deux méthodes de simulation Monte Carlo de matrices de corréla-

tion selon la loi a priori. La première se fonde sur la possibilité de tirages i.i.d. de vecteurs sur un hypersphère, alors que la seconde, elle, repose sur un algorithme de Metropolis. Cette dernière méthode est également utilisée pour la simulation de matrices de corrélation selon la loi a posteriori, qui reflète l'actualisation des croyances initiales sur la structure de corrélation à la lumière de l'observation des données.

Des exercices de simulation illustrent notre méthodologie et permettent de conclure à son efficacité et à son efficience computationnelle.

Remerciements

Ma profonde gratitude va à l'endroit du Professeur William McCausland qui, au-delà de m'avoir initié à l'approche bayésienne de la statistique et ses applications, ainsi qu'à la théorie moderne des probabilités; m'a également mis le pied à l'étrier sur cette recherche stimulante et enrichissante.

Mes remerciements s'étendent également à l'ensemble du Département de sciences économiques de l'Université de Montréal, pour le grand apport intellectuel et l'appui financier dont il m'a permis de bénéficier pendant la durée de mon séjour.

Sommaire

Chapitre1
Introduction

L'analyse statistique des matrices de variance-covariance et, plus singulièrement, des matrices de corrélation continue de donner lieu à une floraison d'initiatives de recherche. En effet, la spécificité de ces objets, en particulier la propriété de définie-positivité qui s'impose à ces matrices, entraîne des considérations particulières d'estimation et d'inférence statistique.

Dans le cas particulier de leur analyse sous le paradigme bayésien, la principale difficulté réside dans la spécification d'une densité a priori. Le rôle de cette densité a priori est de permettre la formulation de croyances de départ sur l'incertitude vis-à-vis du paramètre qui gouverne le processus stochastique en étude; ces croyances a priori devant être raffermies par l'observation de réalisations empiriques (données) du processus. C'est pour cette raison que la plupart des études appliquées font recours aux lois de la famille de Wishart inverse, qui s'accommodent à la propriété définie-positivité tout en permettant, dans le cadre d'un modèle de données gaussien ou conditionnellement gaussien, une densité a posteriori appartenant à la même famille de lois ("conjugaison de la loi de Wishart inverse au modèle gaussien"). Toutefois, cette perspective n'autorise qu'une très faible marge de manoeuvre (à travers les paramètres de la loi Wishart) pour "faire parler" (elicitation) ces lois a priori. Et ce dernier inconvenient constitue une sérieuse entrave au principal atout de l'approche bayésienne par rapport à son homologue fréquentiste: la possibilité d'exprimer des appréhensions initiales à l'égard des paramètres régissant le phénomène étudié.

Le but de ce rapport de recherche est d'apporter une contribution en vue d'atténuer ce dernier désavantage. Nous proposons une façon de spécifier des lois à priori, qui autorise une plus grande marge manoeuvre dans la formulation des croyances initiales sur le degré de corrélation entre différentes variables, tout en respectant les contraintes analytiques des matrices de corrélation. Aussi notre spécification, qui s'inspire volontiers de l'nterprétation géométrique de la notion de corrélation linéaire en termes de produit-scalaire, offre des possibilités d'extension à l'analyse hiérarchique, en intégrant des regroupements de variables en fonction de leurs comportements inter/intra-groupes communs (hiérarchisation).

La démarche d'une modélisation spécifique des matrices de corrélation, indépendamment de l'ensemble de la structure de variance-covariance, se justifie à bien des égards. À titre illustratif, nous donnons ici deux exemples, parmi tant d'autres.

En finance de marché, la famille de modèles ARCH-GARCH bénéficie de beaucoup de notoriété grâce à sa grande capacité de modélisation de la volatilité de nombreux types d'actifs financiers. Cependant, les versions multivariées de ces modèles rencontrent de sérieux problèmes de mise en oeuvre, essentiellement dus à la difficulté de modéliser la dynamique d'une structure (matrice) de variance-covariance de large dimension et qui évolue avec le temps. Pour pallier cette difficulté, Bollerslev (1990) a introduit le modèle GARCH multivarié à corrélations constantes. Cette dernière version du GARCH multivariée modélise séparement les dynamique des volatilités individuelles pour les différents actifs d'une part et, d'autre part, l'ensemble de la structure de corrélation. De ce fait, les volatilités individuelles sont modélisées par des processus dynamques univariés, alors que la structure de

corrélation est supposée invariante par rapport au temps (corrélations constantes). Par cette "double" formulation, le caractère dynamique de l'ensemble de la structure de variance-covariance est préservée (à travers la dynamique des volatilités individuelles des actifs), et la tâche de modélisation de l'ensemble de la structure de variance-covariance est significativement allégée par l'introduction d'une structure multidimensionnelle de corrélations constantes.

De même, en microéconométrie, le modèle probit multivarié est largement utilisé pour la modélisation jointe de plusieurs variables binaires. Dans ce modèle, les probabilités d'occurence des différentes variables dichotomiques sont fonction des effets de plusieurs variables explicatives communes et de chocs corrélés. Cependant, sous les hypothèses statistiques usuelles, l'inférence statistique y est impossible du fait que la matrice de variance-covariance est, au départ, sous-identifiée. Pour permettre l'inférence, il devient donc nécessaire de contraindre cette matrice de façon appropriée, de façon à garantir son identifiabilité. À cet égard, la façon la plus couramment utilisée consiste à réduire les variances individuelles à l'unité, contraignant ainsi la matrice de variance-covariance à être une matrice de corrélation.

Dans la suite du document, l'exposé du travail est organisé comme suit: après un bref rappel des fondements de l'approche bayésienne de l'inférence statistique, nous détaillons la formulation de notre méthodologie et présentons les résultats de son évaluation empirique sur base d'exercices de simulation, pour notre principal modèle dit "de corrélations communes". Par suite, nous proposons une extension de ce dernier pour permettre

une certaine hétérogénéité dans la structure de corrélation, avant de donner l'esquisse d'un modèle plus général qui procède à une hiérarchisation des différentes variables en groupes selon l'homogénéité interne et l'hétérogénéité externe des différents groupes. Nous concluons par des remarques générales.

Chapitre2
Aperçu sur les fondements de l'analyse statistique bayésienne

2.1 Le paradigme bayésien

Jusqu'il y a peu, l'approche la plus répandue d'inférence statistique était celle dite *classique* (encore appelée *fréquentiste* ou *fisherienne)*. En de termes simples, cette approche vise l'estimation et l'inférence sur la valeur certaine mais inconnue d'un paramètre régissant un modèle probabiliste. Cependant, pour arriver à ses fins, cette approche recourt à des procédures qui revêtent de l'incertitude (méthode d'estimation, règle de décision, ...), au-delà de l'incertitude entretenue par les hypothèses de travail statistiques (indépendance des observations, distribution des termes d'erreur, ...).

Dans une optique différente, l'approche *bayésienne*, elle, cherche à formaliser autrement l'incertitude relative aux paramètres inconnus d'un modèle. Elle part d'une formulation de croyances initiales sur les valeurs plausibles de ces paramètres; et fait l'observation des données afin d'en extraire des "signaux" susceptibles d'éclairer et d'orienter les croyances a priori. Du point de vue bayésien donc, une probabilité, au lieu d'être perçue comme la limite d'une fréquence quand le nombre d'expériences se multiplie; elle est plutôt vue comme la traduction numérique du degré de confiance accordé aux différentes hypothèses envisageables sur la valeur d'un paramètre d'intérêt. Ainsi, l'approche bayésienne offre une alternative intéressante d'inférence statistique, en utilisant des méthodes "imperson-

nelles" pour mettre à jour des croyances "personnelles", et offre un cadre élargi d'analyse de l'incertitude sur des paramètres inconnus.

En résumé, il s'agit donc d'un contraste entre l'incertitude des paramètres dans l'approche bayésienne contre l'incertitude des procédures dans l'approche fréquentiste.

Entre autres avantages, l'approche bayesienne requiert moins la disponibilité d'un échantillon important de données, contrairement à l'alternative fréquentiste qui repose largement sur des résultats asymptotiques; et permet également d'isoler de façon efficace et élégante, par intégration, les effets de contamination des paramètres de nuisance.

2.2 La règle de Bayes et l'inférence bayésienne

La règle de Bayes, issue de la théorie des probabilités, constitue la pierre angulaire de l'analyse statistique bayésienne.

Supposons que nous observons une variable aléatoire y, et que nous souhaitons faire de l'inférence sur une autre variable aléatoire θ, les deux variables évoluant selon une certaine distribution de probabilité $p(y, \theta)$, $\theta \in \Theta$.

De la définition de la probabilité conditionnelle, nous avons que:

$$p(\theta \mid y) = \frac{p(y, \theta)}{p(y)}. \tag{2.1}$$

Toujours de cette définition, nous pouvons exprimer la distribution jointe $p(y, \theta)$, en conditionnant par rapport à θ, comme suit:

$$p(y, \theta) = p(\theta \mid y)p(\theta). \tag{2.2}$$

En mettant ensemble les deux précédentes relations, nous obtenons la fondamentale règle de Bayes:

$$p(\theta \mid y) = \frac{p(\theta \mid y)p(\theta)}{p(y)} = \frac{p(y \mid \theta)p(\theta)}{\int_{\Theta} p(y, \theta)d\theta} = \frac{p(y \mid \theta)p(\theta)}{\int_{\Theta} p(y \mid \theta)p(\theta)d\theta}\cdot \tag{2.3}$$

Dans le cadre de l'analyse bayésienne, les objets de la relation (2.3) sont:

- θ, le paramètre d'intérêt (éventuellement multidimensionnel), qui est inconnu ;
- y, les données;
- $p(\theta)$, la distribution a priori, reflétant les croyances originelles sur les valeurs plausibles du paramètre θ;
- $p(\theta \mid y)$, la distribution a posteriori, reflétant les croyances actualisées après l'observation des données y;
- $p(y \mid \theta) \equiv l(\theta \mid y)$, la vraisemblance, encore appelé le "modèle".

Comme la quantité $p(y) = \int_{\Theta} p(y, \theta)d\theta$ est constante (par rapport au paramètre θ), nous avons donc que la distribution a posteriori est proportionnelle au produit de la vraisemblance avec la distribution a priori, ce qui se formule:

$$p(\theta \mid y) \propto l(\theta \mid y)p(\theta)\cdot \tag{2.4}$$

En passant aux logarithmes et en ignorant le terme constant, il suit que:

$$\log\ p(\theta \mid y) = \log p(\theta) + \log\ l(\theta|y)\cdot \tag{2.5}$$

Par conséquent, l'intensité du rapport entre la distribution a posteriori et la distribution a priori (qui peut être évaluée en essayant différentes distributions a priori) fournit une indication de la quantité d'information contenue dans les données, sur les valeurs plausibles du paramètre inconnu. Ainsi, une forte dépendance de la distribution a posteriori sur la distribution a priori pourrait être un indice du faible signal que fournissent les données sur les valeurs du paramètre. À l'opposé, lorsque la distribution a posteriori est faiblement influencée par les changements de distributions a priori, cela présage que les données contiennent beaucoup d'information sur les valeurs plausibles du paramètre.

2.3 Évaluation de la distribution a posteriori

La principale difficulté de l'approche bayésienne réside dans la complication de décrire la distribution a posteriori $p(\theta \mid y)$. En effet, seules quelques formes de distributions a posteriori admettent des formes analytiques usuelles (en particulier, toute loi a priori de la famille exponentielle est conjuguée à certains modèles de données. Autrement dit, étant donné un modèle particulier de données, la distribution a posteriori a la même forme que la distribution a priori. À titre d'exemple, la loi Gamma est conjuguée à un modèle de données de type Poisson.)

Toutefois, cette complication est fortement mitigée par la disponibilité de méthodes efficaces de simulation et d'intégration numériques, connues sous les vocables de méthodes MCMC (Monte Carlo Markov Chain) et IS (Importance Sampling ou Échantillonage d'importance, en français). Dans ce travail, nous allons recourir à l'une d'entre elles,

l'algorithme de Metropolis, qui permet de simuler une distribution a posteriori, afin de procéder à des estimations Monte Carlo.

Le lecteur intéressé peut se référer à Gamerman (2006) pour une vue générale sur les méthodes MCMC et IS.

Chapitre3
Modèle pour une structure de corrélations communes

3.1 Motivation

Le présent chapitre présente une famille de lois a priori pour l'estimation bayésienne de matrices de corrélation. Cette famille de lois se caractérise par le fait que les différentes variables présentent, mutuellement, des degrés de corrélation indépendants et identiquement distribués.

Nous présentons cette méthode sous l'appelation de "modèle pour une structure de corrélations communes", en s'inspirant de la terminologie de Liechty (2004).

3.2 Formulation

Soit C une matrice de corrélation, de format $n \times n$

L'idée-clé de notre méthodologie consiste à formuler nos croyances de base par la spécification d'une loi a priori sur une matrice V, de format $n \times r$, telle que $C = VV^T$. Cette nouvelle matrice V, qui tient lieu de nouvel paramètre d'intérêt, doit satisfaire 3 conditions:

1. $r \geq n$;

2. Toutes les lignes de V sont des vecteurs de norme euclidienne unitaire;

3. V est de plein-rang ligne (son rang est égal à n).

D'une part, les conditions ci-dessus garantissent que la matrice produit VV^T est une matrice symétrique définie-positive, dont les éléments diagonaux sont égaux à l'unité. Par conséquent, la matrice VV^T est une matrice de corrélation.

D'autre part, pour toute matrice de corrélation C, il existe toujours au moins une matrice V qui satisfait aux conditions précitées, et telle que $VV^T = C$. Pour s'en convaincre, considérons la décomposition de Cholesky de la matrice C. Rappelons que pour toute matrice symétrique définie-positive C, la décomposition de Cholesky donne une matrice triangulaire inférieure L telle que $C = LL^T$. En prenant $V = [L, 0_{n,r-n}]$, nous obtenons donc une matrice V qui satisfait aux conditions requises.

Ainsi, une densité a priori sur V induit implicitement une densité a priori sur la matrice de corrélation C.

Notons par ailleurs qu'une telle formulation se prête bien aux interprétations géométriques: en effet, une matrice V, telle que spécifiée, représente une séquence de n points $(r-uplets)$ sur l'hypersphère de dimension r, centrée à l'origine et de rayon unité.

3.3 Spécification

Soit le vecteur $\bar{v}_r \equiv (1, 0, 0, ..., 0)$ de $\mathbb{R}^r$, dénommé "pôle nord" du *r-hypersphère standard* dans la suite du texte;

Soit C une matrice de corrélation donnée;

Soit V, une matrice $n \times r$ telle que définie dans la section précédente;

Soit v_i, la i-ème ligne de V;

Les vecteurs $v_i, i = 1\ 2, ..., n$, lignes de la matrice V, sont indépendamment et identiquement distribués (i.i.d);

Soit θ_i, l'angle que forme le vecteur v_i avec le pole nord $\bar{v}_r$;

Nous formulons notre loi a priori, en deux temps, de la manière suivante:

- Nous spécifions d'abord une loi marginale $f(\theta_i)$ pour l'angle θ_i. Par exemple, on peut choisir $f(\theta_i)$ dans la famille de lois de probabilité Bêta, qui ont l'avantage de présenter un support de probabilité continu et fini, l'intervalle $[0; 1]$, et en l'étendant homothétiquement sur l'intervalle $[0; \pi]$, support défini pour θ_i;

- Conditionnellement à l'angle θ_i, le vecteur v_i suit une loi uniforme sur l'ensemble des points de l'hypersphère de rayon unitaire qui forme un angle θ_i avec le pole nord. Cet ensemble correspond à la surface d'un hypersphère de dimension $(r-1)$ et de rayon $\sin\theta_i$;

Il s'agit donc d'une densité, qui s'écrit:

$$f(v_i) = f(\theta_i).\left[\frac{(2\pi)^{(r-1)/2}}{\Gamma(\frac{r-1}{2})}.\sin^{r-2}\theta_i\right]^{-1}; \tag{3.6}$$

où $\theta_i = \cos^{-1}(v_i.\bar{v}_r) = \cos^{-1}(v_{i1})$ est l'angle entre les vecteurs v_i et $\bar{v}_r$.

3.4 Simulation selon la loi a priori par la méthode directe (tirages i.i.d.)

Motivation

Dans notre modèle, comme dans un grand nombre de cas d'analyse bayesienne, la simulation de matrices de corrélation selon la loi a posteriori ne peut pas être faite d'une manière directe. Elle requiert impérativement le recours à une méthode MCMC, l'algorithme de Metropolis.

À l'opposé, la simulation selon la loi a priori, elle, offre un choix: elle peut se faire soit par une méthode directe, soit par une méthode indirecte à travers l'algorithme de Metropolis. Pour cette raison, afin de pouvoir évaluer l'efficacité de l'algorithme de Metropolis que nous allons définir, il est intéressant de comparer les résultats des deux méthodes de simulation selon la loi a priori (méthode directe et méthode Metropolis). La proximité des résultats fournis par les deux méthodes de la simulation a priori serait alors une indication de l'efficacité de la méthode Metropolis, et conforterait donc son utilisation de cette dernière pour la simulation a posteriori pour laquelle la méthode directe est inopérante.

La sous-section suivante détaille les étapes de mise en oeuvre de la méthode directe.

Mise en oeuvre

Le tirage d'un vecteur v_i selon la densité à priori se fait en deux étapes, précisées ci-après:

- D'abord, on tire l'angle θ_i de sa loi marginale. Il s'agit ici d'un simple tirage univarié;

- Ensuite, le tirage du vecteur v_i de sa distribution conditionnelle sachant θ_i se fait en tirant un vecteur de la distribution uniforme sur l'hypersphère de dimension $(r-1)$ et de rayon $\sin\theta_i$. Ceci peut notamment se faire en tirant un vecteur aléatoire d de loi normale multivariée matrice de variance-covariance identitaire, dont la longueur est normalisée à l'unité.

  Les j coordonnées de v_i, $j = 1$ *à* r, sont alors données par la relation:

$$v_i = (\cos\theta_i, d_1 \sin\theta_i, d_2 \sin\theta_i, ..., d_{r-1} \sin\theta_i) \cdot \tag{3.7}$$

Étant donné le vecteur d'hyperparamètres $\theta = (\theta_1, \theta_2, ..., \theta_n)$, l'espérance conditionnelle et la limite en probabilité de la matrice de corrélation $C = V.V^T$ sont données par:

$$\text{plim}_{r\to\infty} C \mid \theta = E[C \mid \theta] = \begin{bmatrix} \cos\theta_1 \\ \cdot \\ \cdot \\ \cos\theta_n \end{bmatrix} . [\cos\theta_1, ..., \cos\theta_n] + \begin{bmatrix} \sin^2\theta_1 & & 0 \\ & \cdot & \\ & \cdot & \\ 0 & & \sin^2\theta_n \end{bmatrix} .\cdot$$

Cette dernière relation présente un double intérêt:

- D'une part, dans le cas où $r = \infty$, l'angle formé par deux vecteurs uniformément distribués sur un même hypersphère vaut $\frac{\pi}{2}$ avec probabilité 1. En conséquence, la corrélation entre une paire de vecteurs quelconques est complétement déterminée par la donnée des angles d'orientation respectifs θ_i de ces vecteurs par rapport à l'axe du pôle nord. Cette "marginalisation" autorise donc de pouvoir travailler uniquement avec les angles d'orientation θ_i, en lieu et place des vecteurs v_i en entier;

- D'autre part, au sein de certains ensembles de variables qui présentent une structure complexe, le comportement des vecteurs v_i donnés par une factorisation du type

$C = VV^T$ pourrait ne pas être invariant à l'ordonnancement des différentes variables dans la structure de corrélation. Ceci serait alors de nature à poser des problèmes de cohérence d'analyse, car les résultats obtenus dépendraient de l'ordre arbitraire initialement attribué aux variables. En apportant une symétrie complète entre les variables, dont le comportement ne dépend alors que des angles d'orientation θ_i respectifs, la relation ci-haut permet d'éviter une dépendance éventuelle des résultats obtenus, à l'ordre arbitraire attribué aux différentes variables.

3.5 Simulation selon la loi a priori par un algorithme de Metropolis

Conformément à la motivation donnée en début de la section précédente (simulation selon la loi a priori par la méthode directe), une méthode alternative de simulation de matrices de corrélation selon la densité a priori est la version suivante de l'algorithme de Metropolis dont le noyau (mécanisme de proposition) suit une marche aléatoire. Par cet algorithme, l'angle θ qui gouverne la marche a une distribution aléatoire arbitraire mais la direction de la marche d'un vecteur v_i sur l'hypersphère est uniforme.

Les différentes étapes de simulation de la châine markovienne de l'algorithme sont les suivantes:

1. D'abord, nous tirons θ d'une variable aléatoire quelconque, judicieusement définie, ayant pour support l'intervalle $[0, \overline{\theta}]$, où: $0 < \overline{\theta} \leq \pi$. $\overline{\theta}$ est alors un paramètre d'"ajustement" de l'incrément du tirage;

2. Ensuite, nous tirons un vecteur d suivant la loi uniforme sur l'hypersphère de dimension r et de rayon unitaire suivant le procédé expliqué à la section 3.4 précédente;

3. Par suite, nous calculons le vecteur $d_\perp$, qui est la projection de d sur l'hyperplan perpendiculaire au vecteur observé v_i. Ainsi, $d_\perp = d - \frac{v_i.d}{\|v_i\|^2}.v_i$;

4. Finalement, le vecteur instrumental (ou proposition) de cette étape est défini comme: $v_i^* = \cos(\theta_i).v_i + \sin(\theta_i).d_\perp/\|d_\perp\|$. À chaque étape, l'actualisation d'une ligne v_i se fait de façon suivante:

   - Nous déterminons d'abord le ratio $l(\theta_i, \theta_i^*) = \frac{f(v_i^*)}{f(v_i)} = \frac{f(\theta_i^*).\left[\frac{(2\pi)^{(r-1)/2}}{\Gamma(\frac{r-1}{2})}.\sin^{r-2}\theta_i^*\right]^{-1}}{f(\theta_i).\left[\frac{(2\pi)^{(r-1)/2}}{\Gamma(\frac{r-1}{2})}.\sin^{r-2}\theta_i\right]^{-1}}$;

     où $\begin{cases} \theta_i^* = \arccos(v_{i1}^*) \\ \theta_i = \arccos(v_{i1}) \end{cases}$;

   - La probabilité d'acceptation du candidat v_i^* est $\alpha(\theta_i, \theta_i^*) = \min(l(\theta_i, \theta_i^*), 1)$;

   - On tire une valeur $\overline{u}$ selon une variable aléatoire uniforme définie sur l'intervalle $[0; 1]$; et la règle d'acceptation du candidat proposé est la suivante:

     Si $\begin{cases} \alpha(\theta_i, \theta_i^*) \geq \overline{u} & \text{alors le candidat } v_i^* \text{ est accepté et remplace le vecteur actuel } v_i; \\ \alpha(\theta_i, \theta_i^*) < \overline{u} & \text{alors le candidat } v_i^* \text{ est rejeté et le vecteur actuel } v_i \text{ est maintenu;} \end{cases}$;

   - Après chaque étape (actualisation de toutes les lignes), on calcule la matrice de corrélation $C = VV^T$ correspondante.

Cette procédure est répétée un nombre M d'itérations suffisamment grand (par défaut, nous retenons $M = 20000$ au cours du présent travail, à moins qu'une autre valeur

ne soit explicitement déclarée) pour s'assurer qu'après quelques itérations d'initialisation, les matrices V obtenues suivent la loi a priori des lignes v_i. Par conséquent, en vertu d'une loi des grands nombre, nous avons que la moyenne empirique des matrices de corrélation $C = VV^T$ ainsi obtenues, converge presque sûrement vers l'espérance de la matrice de corrélation selon la densité a priori, quand le nombre M des itérations tend vers l'infini.

3.6 Simulation selon la loi a posteriori par un algorithme de Metropolis

La simulation de matrices de corrélation suivant la densité a posteriori suit les grandes lignes de l'algorithme de Metropolis décrit ci-dessus.

Dans le cadre du présent travail, nous considérons une suite d'observations indépendantes et identiquement distribuées $y_t \sim N(0; Q)$, $t = 1, 2, ..., T$. Il s'agit donc d'une restriction pour laquelle la matrice de variance-covariance Q est une matrice de corrélation. À titre d'exemple, une telle spécification peut servir dans le cadre d'une estimation de la matrice de corrélation d'un modèle GARCH à corrélations constantes, tel que mentionné précédemment.

Dans ce cas, la vraisemblance du modèle s'écrit:

$$l(Q \mid \{y_t\}_{1 \leq t \leq T}) = \frac{|Q|^{-\frac{T}{2}}}{(2\pi)^{\frac{nT}{2}}} \exp\{-\frac{1}{2} \sum_{t=1}^{T} y_t' \, Q^{-1} \, y_t\} \cdot \tag{3.8}$$

La seule modification par rapport à l'algorithme précédent de simulation selon la loi a priori consiste donc à intégrer la vraisemblance du modèle dans le processus de proposition de la chaîne markovienne. Le critère (ratio) de décision devient alors:

$$l(\theta_i, \theta_i^*) = \frac{f(v_i^*)}{f(v_i)} = \frac{\{\frac{|Q^*|^{-\frac{T}{2}}}{(2\pi)^{\frac{nT}{2}}} \exp(-\frac{1}{2}\sum_{t=1}^{T} y_t' \, Q^{*-1} \, y_t)\}\{f(\theta_i^*)\left[\frac{(2\pi)^{(r-1)/2}}{\Gamma(\frac{r-1}{2})}.\sin^{r-2}\theta_i^*\right]^{-1}\}}{\{\frac{|Q|^{-\frac{T}{2}}}{(2\pi)^{\frac{nT}{2}}} \exp\{-\frac{1}{2}\sum_{t=1}^{T} y_t' \, Q^{-1} \, y_t\}\}\{f(\theta_i).\left[\frac{(2\pi)^{(r-1)/2}}{\Gamma(\frac{r-1}{2})}.\sin^{r-2}\theta_i\right]^{-1}\}};$$

$$\text{où} \begin{cases} \theta_i^* = \arccos(v_{i1}^*) \text{ et } Q^* \text{ est la matrice obtenue en considerant le vecteur proposé } v_i^* \\ \theta_i = \arccos(v_{i1}) \text{ et } Q \text{ est la matrice obtenue en considérant le vecteur initial } v_i \end{cases}.$$

3.7 Exemple 1: Illustration d'un cas de variables décorrélées

Pour illustrer la performance de notre modèle de corrélations communes, nous allons reprendre dans cette section les résultats d'un exemple de simulation de matrices de corrélations pour lesquelles toute paire de variables présente une corrélation d'espérance nulle (variables décorrélées). L'hyperparamètre θ est simulé suivant une loi Bêta standard (ayant pour support l'intervalle $[0; 1]$), dont les réalisations seront ramenées de façon homothétique à l'échelle appropriée (intervalle $[0, \pi]$).

Les paramètres de simulation sont: $n = 6$; $r = 6$; $\theta/\pi \sim B\hat{e}ta(1; 1)$ $(Loi\ uniforme\ sur\ [0; 1]) \Rightarrow E(\theta) = \frac{\pi}{2}$.

Dans ce cas, les vecteurs v_i sont donc distribués, en moyenne, sur l'hyperplan orthogonal à l'axe du pôle nord. D'où une corrélation de moyenne nulle entre les variables respectives. Dans les tableaux suivants, nous reprenons les moyennes empiriques de la matrice de corrélation C et des angles d'orientation θ, ainsi que les écarts-type respectifs, notés entre parenthèses en dessous des valeurs moyennes des paramètres auxquels ils se réfèrent.

Simulation a priori de matrices de corrélation: méthode i.i.d

Espérance empirique

Matrice de corrélation						θ
1.0000	-0.0000	0.0029	-0.0031	-0.0043	-0.0026	1.5736
(0.0000)	(0.5423)	(0.5414)	(0.5379)	(0.5385)	(0.5400)	(0.9086)
-0.0000	1.0000	-0.0072	0.0003	0.0010	-0.0006	1.5553
(0.5423)	(0.0000)	(0.5396)	(0.5397)	(0.5403)	(0.5406)	(0.9079)
0.0029	-0.0072	1.0000	-0.0012	0.0052	0.0004	1.5796
(0.5414)	(0.5396)	(0.0000)	(0.5386)	(0.5369)	(0.5404)	(0.9052)
-0.0031	0.0003	-0.0012	1.0000	0.0085	-0.0029	1.5708
(0.5379)	(0.5397)	(0.5386)	(0.0000)	(0.5397)	(0.5395)	(0.9041)
-0.0043	0.0010	0.0052	0.0085	1.0000	0.0032	1.5626
(0.5385)	(0.5403)	(0.5369)	(0.5397)	(0.0000)	(0.5385)	(0.9037)
-0.0026	-0.0006	0.0004	-0.0029	0.0032	1.0000	1.5679
(0.5400)	(0.5406)	(0.5404)	(0.5395)	(0.5385)	(0.0000)	(0.9073)

Simulation a priori de matrices de corrélation: algorithme de Metropolis

Espérance empirique

Matrice de corrélation						θ
1.0000	-0.0349	-0.0075	-0.0949	0.0135	-0.0296	1.4668
(0.0000)	(0.5516)	(0.5567)	(0.5894)	(0.5497)	(0.5592)	(0.9086)
-0.0349	1.0000	0.0033	0.0084	0.0416	-0.0461	1.4494
(0.5516)	(0.0000)	(0.5415)	(0.5436)	(0.5304)	(0.5422)	(0.9079)
-0.0075	0.0033	1.0000	-0.0499	-0.0916	0.0354	1.7163
(0.5567)	(0.5415)	(0.0000)	(0.5444)	(0.5264)	(0.5401)	(0.9052)
-0.0949	0.0084	-0.0499	1.0000	-0.0082	-0.0541	1.7756
(0.5894)	(0.5436)	(0.5444)	(0.0000)	(0.5268)	(0.5619)	(0.9041)
0.0135	0.0416	-0.0916	-0.0082	1.0000	-0.0277	1.4863
(0.5497)	(0.5304)	(0.5264)	(0.5268)	(0.0000)	(0.5123)	(0.9037)
-0.0296	-0.0461	0.0354	-0.0541	-0.0277	1.0000	1.5837
(0.5592)	(0.5422)	(0.5401)	(0.5619)	(0.5123)	(0.0000)	(0.9073)

En guise d'évaluation de l'efficacité de notre methodologie pour la simulation a posteriori, nous générons une séquence i.i.d de $T = 150$ vecteurs d'observations y_t suivant la loi normale multivariée $N(0; Q)$, où Q est une matrice de corrélation d'ordre $n = 6$, présentant une structure de corrélation hétérogène entre les différentes variables. Q est fixé pour les différents exercices de simulation tout au long du présent travail.

Q, matrice de corrélation servant de paramètre pour le modèle de données

$y_t \sim N(0; Q), t = 1 \grave{a}\, T$

Matrice de corrélation

1	0.9	0.1	0	-0.5	-0.6
0.9	1	-0.1	-0.2	-0.6	-0.7
0.1	-0.1	1	0.6	0.2	0.1
0	-0.2	0.6	1	0	-0.1
-0.5	-0.6	0.2	0	1	0.3
-0.6	-0.7	0.1	-0.1	0.3	1

Globalement, la structure de corrélation décrite par cette matrice fait ressortir 3 paires de variables ({1,2};{3,4};{5,6}) dont les variables appartennant à chaque paire sont proches, en termes de corrélation, vis-à-vis des autres paires.

Notre but est alors de recouvrer, en termes de moyenne empirique, la matrice de variance-covariance Q (qui est une matrice de corrélation) par le biais d'une simulation a posteriori, telle que décrite dans la section 3.6 , étant donné la spécification de la loi a priori ci-haut mentionnées.

Simulation a posteriori de matrices de corrélation

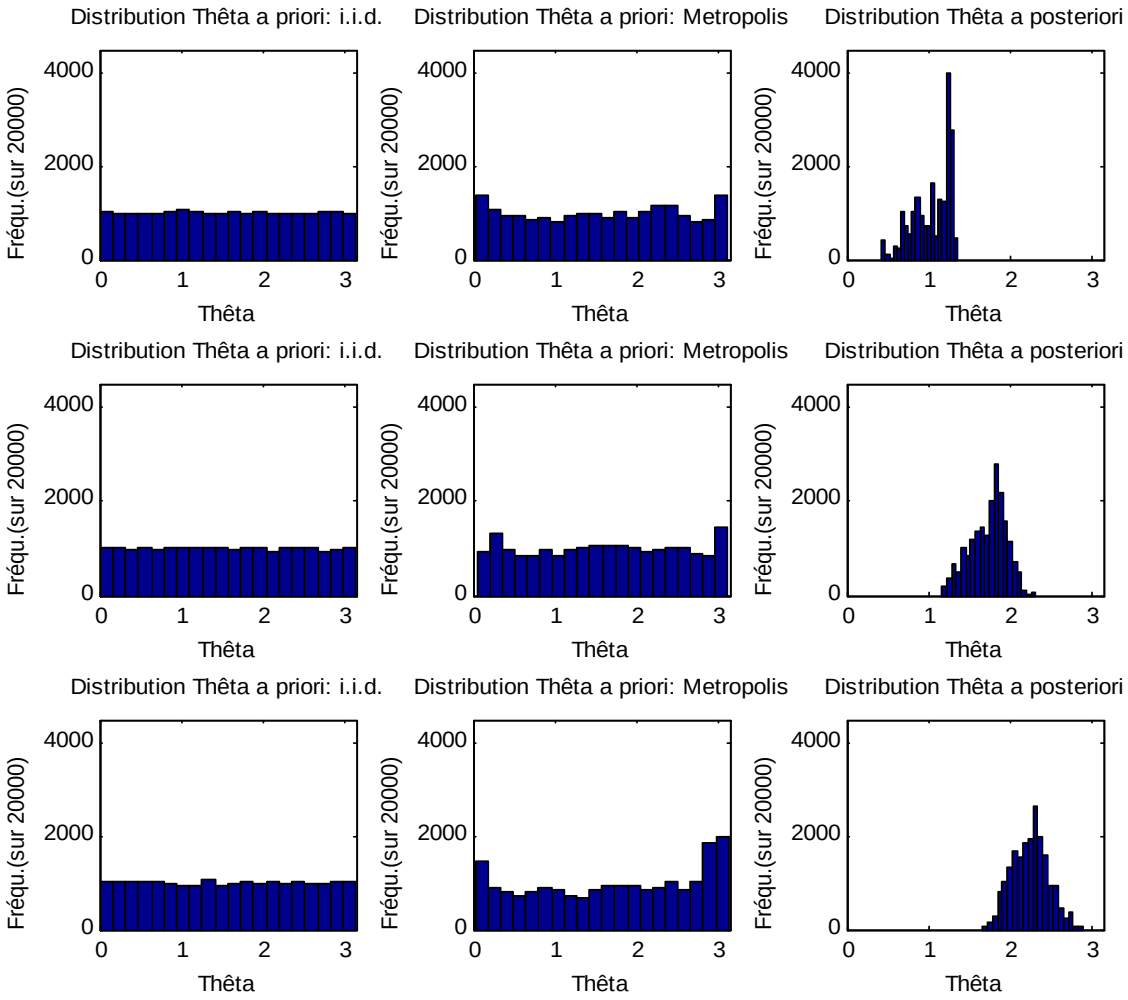
Espérance empirique

Matrice de corrélation						θ
1.0000	0.9005	0.1729	0.0518	-0.5187	-0.5838	1.9796
(0.0000)	(0.0188)	(0.0627)	(0.0567)	(0.0486)	(0.0367)	(0.9086)
0.9005	1.0000	0.0184	-0.1417	-0.5686	-0.6785	1.9097
(0.0188)	(0.0000)	(0.0644)	(0.0569)	(0.0480)	(0.0356)	(0.9079)
0.1729	0.0184	1.0000	0.5299	0.1215	0.0618	2.1226
(0.0627)	(0.0644)	(0.0000)	(0.0440)	(0.0654)	(0.0632)	(0.9052)
0.0518	-0.1417	0.5299	1.0000	-0.1245	-0.1087	2.1818
(0.0567)	(0.0569)	(0.0440)	(0.0000)	(0.0621)	(0.0624)	(0.9041)
-0.5187	-0.5686	0.1215	-0.1245	1.0000	0.2301	1.3797
(0.0486)	(0.0480)	(0.0654)	(0.0621)	(0.0000)	(0.0602)	(0.9037)
-0.5838	-0.6785	0.0618	-0.1087	0.2301	1.0000	1.1463
(0.0367)	(0.0356)	(0.0632)	(0.0624)	(0.0602)	(0.0000)	(0.9073)

Le graphique suivant donne, pour chacune des 3 paires de variables mentionnées ci-dessus, la distribution empirique a priori de l'angle d'orientation θ selon la méthode directe (tirages i.i.d.) et selon la méthode de l'algorithme de Metropolis, de même que sa distribution empirique a posteriori (obtenue également par la méthode de Metropolis).

Distributions a priori (i.i.d et Metropolis) et a posteriori de l'angle d'orientation θ,

pour les différents groupes de variables



Des résultats ci-dessus, nous constatons que l'espérance mathématique de la simulation a posteriori est très proche de Q, le paramètre matriciel qui sous-tend le modèle des données. La méthodologie élaborée parvient ainsi à recouvrer efficacement le paramètre Q d'intérêt, en termes de moyenne. De plus, en partant d'une loi a priori très diffuse (loi uniforme) de θ , les distributions a posteriori obtenues présentent des formes concentrées autour de certaines valeurs du paramètre θ.

3.8 Exemple 2: Évaluation de la robustesse face aux différences de spécification de la loi a priori

Dans le but de se faire une idée sur la sensibilité des résultats de notre méthodologie de simulation a posteriori, face à de notables différences entre les lois a priori considérées, nous allons reprendre les 3 étapes de la section précédente (simulation a priori selon la méthode i.i.d; simulation a priori selon un algorithme de Metropolis, et simulation a posteriori). Les nouvelles simulations utilisent un nouvel ensemble de paramètres, différent de celui du cas précédent. Mais les observations, elles, suivent toujours notre modèle de données ($y_t \sim N(0; Q), t = 1, 2, ..., T.$)

Ainsi, l'objectif visé est d'évaluer dans quelle mesure on retrouve la matrice de variance-covariance qui sous-tend nos observations (la matrice de corrélation Q), malgré une spécification de la loi a priori qui s'écarte sensiblement de la spécification précédente. En cas de succès, cela serait une indication de la robustesse de la loi a posteriori.par rapport aux différentes spécifications de loi a priori.

Les nouveaux paramètres de simulation sont donc: $n = 6$; $r = 7$; $\theta/\pi \sim B\hat{e}ta(1; 9) \Rightarrow E(\theta) = \frac{\pi}{10}$; $T = 150$. Notons, au passage, que cette dernière spécification donne un modèle de matrices qui présentent des corrélations identiquement et fortement et corrélées, en moyenne.

Simulation a priori de matrices de corrélation: méthode i.i.d.

Espérance mathématique

Matrice de corrélation						θ
1.0000	0.8383	0.8374	0.8352	0.8360	0.8374	0.3146
(0.000)	(0.1944)	(0.1964)	(0.1976)	(0.1969)	(0.1961)	(0.2866)
0.8383	1.0000	0.8394	0.8371	0.8372	0.8381	0.3137
(0.1944)	(0.000)	(0.1944)	(0.1964)	(0.1950)	(0.1942)	(0.2838)
0.8374	0.8394	1.0000	0.8373	0.8378	0.8372	0.3136
(0.1964)	(0.1944)	(0.000)	(0.1955)	(0.1966)	(0.1972)	(0.2864)
0.8352	0.8371	0.8373	1.0000	0.8354	0.8365	0.3172
(0.1976)	(0.1964)	(0.1955)	(0.000)	(0.1988)	(0.1961)	(0.2878)
0.8360	0.8372	0.8378	0.8354	1.0000	0.8373	0.3162
(0.1969)	(0.1950)	(0.1966)	(0.1988)	(0.000)	(0.1966)	(0.2874)
0.8374	0.8381	0.8372	0.8365	0.8373	1.0000	0.3144
(0.1961)	(0.1942)	(0.1972)	(0.1961)	(0.1966)	(0.000)	(0.2860)

Simulation a priori de matrices de corrélation: algorithme de Metropolis

Espérance empirique

Matrice de corrélation						θ
1.0000	0.8345	0.8002	0.8102	0.8538	0.8202	0.3641
(0.000)	(0.1820)	(0.1987)	(0.1976)	(0.1526)	(0.1969)	(0.2866)
0.8345	1.0000	0.8234	0.8292	0.8714	0.8422	0.3055
(0.1820)	(0.000)	(0.1865)	(0.1915)	(0.1560)	(0.1946)	(0.2838)
0.8002	0.8234	1.0000	0.7990	0.8417	0.8116	0.3791
(0.1987)	(0.1865)	(0.000)	(0.2169)	(0.1800)	(0.2031)	(0.2864)
0.8102	0.8292	0.7990	1.0000	0.8487	0.8175	0.3617
(0.1976)	(0.1915)	(0.2169)	(0.000)	(0.1823)	(0.2099)	(0.2878)
0.8538	0.8714	0.8417	0.8487	1.0000	0.8598	0.2500
(0.1526)	(0.1560)	(0.1800)	(0.1823)	(0.000)	(0.1716)	(0.2874)
0.8202	0.8422	0.8116	0.8175	0.8598	1.0000	0.3380
(0.1969)	(0.1946)	(0.2031)	(0.2099)	(0.1716)	(0.000)	(0.2860)

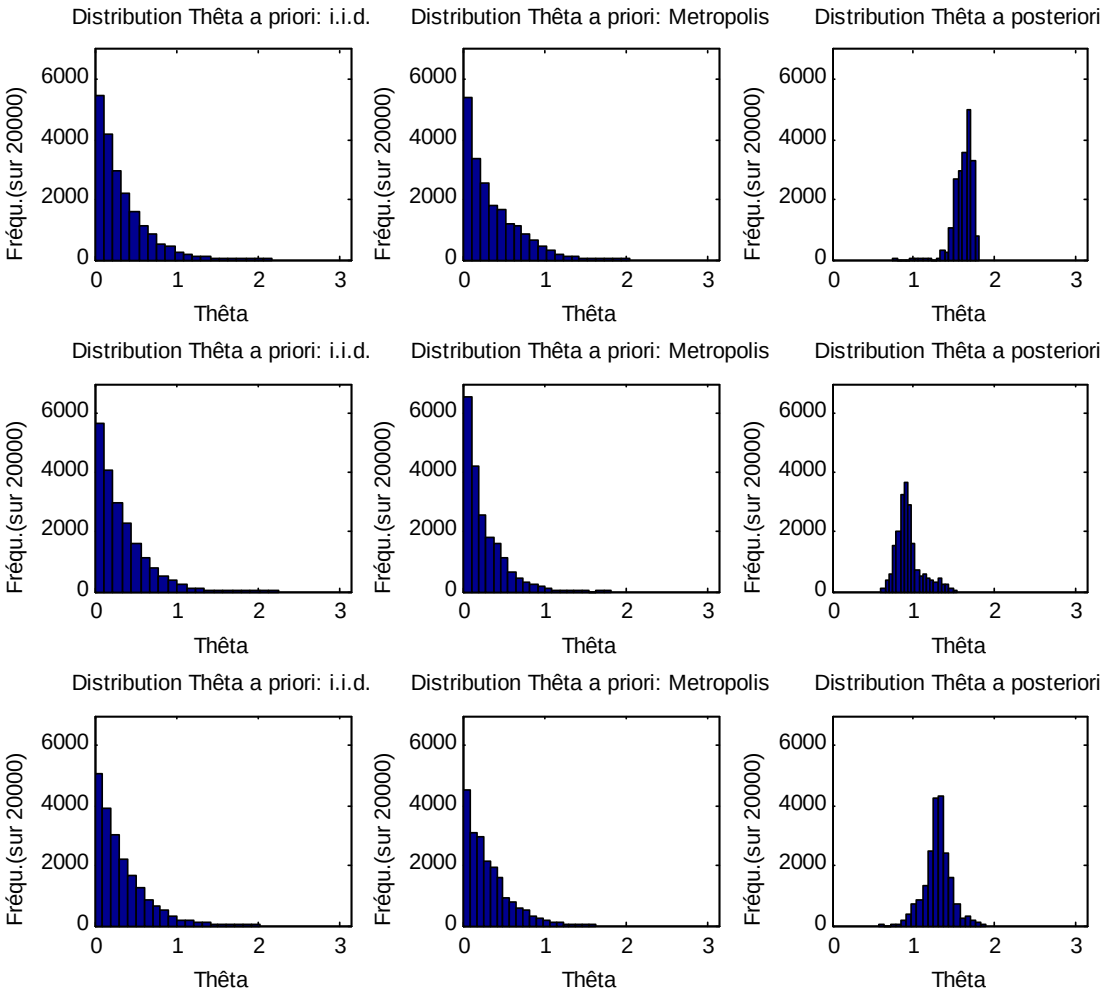
Simulation a posteriori de matrices de corrélation

Espérance empirique

Matrice de corrélation						θ
1.0000	0.8744	0.1393	-0.0214	-0.4766	-0.5676	1.6293
(0.000)	(0.0136)	(0.0655)	(0.0604)	(0.0499)	(0.0511)	(0.2866)
0.8744	1.0000	-0.1002	-0.2367	-0.5895	-0.6901	1.8878
(0.0136)	(0.000)	(0.0659)	(0.0625)	(0.0472)	(0.0427)	(0.2838)
0.1393	-0.1002	1.0000	0.6521	0.1666	0.0289	0.7571
(0.0655)	(0.0659)	(0.000)	(0.0381)	(0.0634)	(0.0697)	(0.2864)
-0.0214	-0.2367	0.6521	1.0000	-0.0040	-0.0911	0.5978
(0.0604)	(0.0625)	(0.0381)	(0.000)	(0.0634)	(0.0637)	(0.2878)
-0.4766	-0.5895	0.1666	-0.0040	1.0000	0.2715	1.4100
(0.0499)	(0.0472)	(0.0634)	(0.0634)	(0.000)	(0.0592)	(0.2874)
-0.5676	-0.6901	0.0289	-0.0911	0.2715	1.0000	1.4375
(0.0511)	(0.0427)	(0.0697)	(0.0637)	(0.0592)	(0.000)	(0.2860)

Distributions a priori (i.i.d et Metropolis) et a posteriori de l'angle d'orientation θ,

pour les différents groupes de variables



Les résultats ci-dessus montrent que, malgré une différence sensible dans la spécification de la loi a priori par rapport au cas précédent, la méthodologie développée parvient à retrouver (en termes de moyenne) la matrice qui sous-tend notre modèle d'observations. Aussi, l'analyse graphique montre qu'à partir d'une distribution fortement asymétrique et diffuse de la loi a priori de l'hyperparamètre θ (d'espérance $\frac{\pi}{10}$), la simulation a posteriori nous donne des distributions symétriques et concentrées autour de certaines valeurs de θ.

Il s'agit donc clairement d'une indication de la robustesse de notre modèle face à des différences significatives dans la spécification de la loi a priori.

3.9 Limites

La principale limite de notre modèle est sa capacité de modélisation réduite. En effet, le modèle à corrélations communes est très restrictif et se prête peu aux matrices de corrélation couramment observées. En effet, l'intensité de la liaison linéaire entre plusieurs variables, prises 2 à 2, varie substantiellement en fonction du lien intrinsèque qui existe entre elles. Par conséquent, le modèle développé dans ce chapitre, qui suppose une évolution linéaire a priori identique entre toutes les variables, due à notre hypothèse de départ sur l'identité des distributions $f(\theta)$ est, dans la pratique, d'une portée fort limitée.

Afin de remédier, quoique très partiellement, à cet inconvénient, le chapitre suivant développe une extension du modèle précédent, sous l'appellation de "modèle de corrélations groupées", dans la droite ligne de la terminologie de Liechty (2004).

Chapitre4
Extensions

4.1 Modèle de corrélations groupées

4.1.1 Motivation

L'extension du modèle précédent, que nous appelons "modèle de corrélations groupées", remédie dans une étroite mesure, aux insuffisances du modèle originel (modèle à corrélations communes). Cette extension simple et directe permet d'apporter une certaine diversité dans la structure de corrélations a priori d'un ensemble de variables

4.1.2 Formulation, Spécification et Simulation

La formulation, la spécification ainsi que les simulations a priori et a posteriori sont similaires à celles du modèle de base. Le seul changement apporté ici est la relaxation de l'hypothèse de distributions *identiques* des vecteurs-lignes v_i du paramètre matriciel V. La différence dans la distribution des v_i est apportée par la différence de distribution des angles d'orientation θ_i, tout en maintenant, pour toutes les vecteurs v_i, une distribution conditionnelle uniforme sur l'hypersphère de dimension $(r-1)$, une fois l'angle d'orientation θ fixé. Ce faisant, la dissemblance d'orientation des vecteurs v_i par rapport à l'axe du pôle nord apporte de la diversité dans l'ensemble de la structure de corrélation a priori.

Pour un vecteur v_i donné, i^{ème} ligne du paramètre matriciel V, la densité a priori s'écrit toujours selon la formule (3.6), à la seule différence que pour le cas présent, les angles θ_i sont indépendamment mais *non-identiquemment* distribués.

4.1.3 Illustration

Nous illustrons notre modèle par des simulations a priori et a posteriori pour 3 groupes de variables, suivant le jeu de paramètres suivant:

Groupe 1: $n_1 = 2$, $\theta_1/\pi \sim B\hat{e}ta(1; 5)$;

Groupe 2: $n_2 = 2$, $\theta_2/\pi \sim B\hat{e}ta(5; 5)$;

Groupe 3: $n_3 = 2$, $\theta_3/\pi \sim B\hat{e}ta(5; 1)$;

On a donc $n = n_1 + n_2 + n_3 = 6$; $r = 7$

Simulation a priori de matrices de corrélation: méthode i.i.d.

Espérance empirique

Matrice de corrélation						θ
1.0000	0.6332	-0.0039	0.0003	-0.6292	-0.6256	0.5229
(0.0000)	(0.3539)	(0.4148)	(0.4132)	(0.3563)	(0.3616)	(0.4421)
0.6332	1.0000	-0.0036	-0.0009	-0.6287	-0.6265	0.5228
(0.3539)	(0.0000)	(0.4163)	(0.4140)	(0.3559)	(0.3585)	(0.4401)
-0.0039	-0.0036	1.0000	-0.0018	0.0038	0.0051	1.5757
(0.4148)	(0.4163)	(0.0000)	(0.3798)	(0.4139)	(0.4151)	(0.4738)
0.0003	-0.0009	-0.0018	1.0000	-0.0003	-0.0011	1.5721
(0.4132)	(0.4140)	(0.3798)	(0.0000)	(0.4114)	(0.4131)	(0.4718)
-0.6292	-0.6287	0.0038	-0.0003	1.0000	0.6208	2.6121
(0.3563)	(0.3559)	(0.4139)	(0.4114)	(0.0000)	(0.3629)	(0.4469)
-0.6256	-0.6265	0.0051	-0.0011	0.6208	1.0000	2.6074
(0.3616)	(0.3585)	(0.4151)	(0.4131)	(0.3629)	(0.0000)	(0.4511)

Simulation a priori de matrices de corrélation: algorithme de Metropolis

Espérance empirique

Matrice de corrélation						θ
1.0000	0.6112	0.0034	-0.0026	-0.6875	-0.6214	0.4903
(0.0000)	(0.3905)	(0.4093)	(0.4204)	(0.3432)	(0.3732)	(0.4421)
0.6112	1.0000	0.0038	-0.0074	-0.6559	-0.5864	0.5689
(0.3905)	(0.0000)	(0.4111)	(0.4102)	(0.3618)	(0.3878)	(0.4401)
0.0034	0.0038	1.0000	0.0052	-0.0157	-0.0160	1.5586
(0.4093)	(0.4111)	(0.0000)	(0.3732)	(0.4144)	(0.4120)	(0.4738)
-0.0026	-0.0074	0.0052	1.0000	0.0088	0.0118	1.5772
(0.4204)	(0.4102)	(0.3732)	(0.0000)	(0.4141)	(0.4076)	(0.4718)
-0.6875	-0.6559	-0.0157	0.0088	1.0000	0.6659	2.7367
(0.3432)	(0.3618)	(0.4144)	(0.4141)	(0.0000)	(0.3495)	(0.4469)
-0.6214	-0.5864	-0.0160	0.0118	0.6659	1.0000	2.5759
(0.3732)	(0.3878)	(0.4120)	(0.4076)	(0.3495)	(0.0000)	(0.4511)

Pour la simulation a posteriori, nous générons une séquence i.i.d de $T = 150$ vecteurs d'observations suivant notre modèle de données $y_t \sim N(0; Q), t = 1, 2, ..., T$.

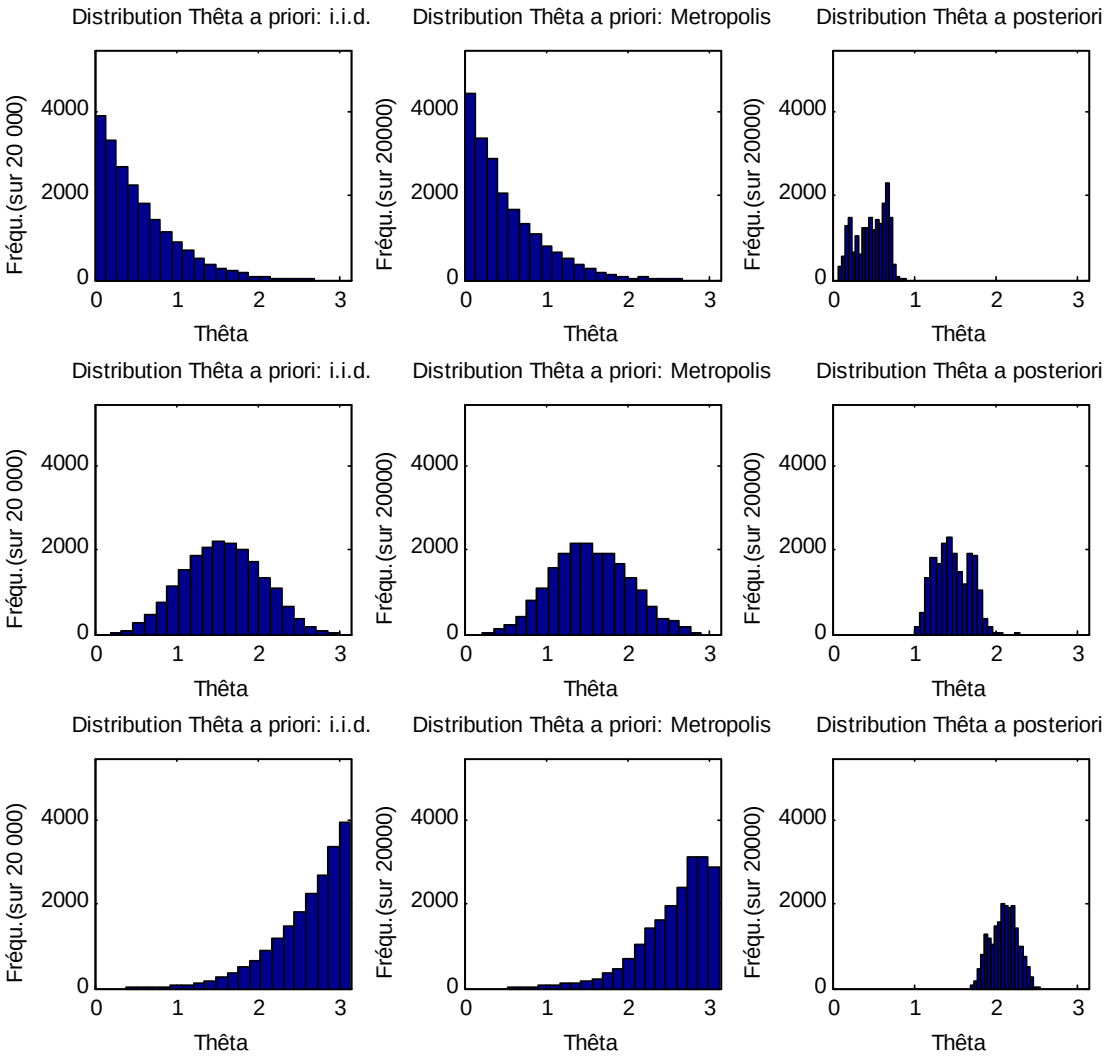
Simulation a posteriori de matrices de corrélation

Espérance empirique

Matrice de corrélation						θ
1.0000	0.9111	0.0644	0.0289	-0.5209	-0.6192	0.8765
(0.0000)	(0.0257)	(0.0595)	(0.0517)	(0.0509)	(0.0400)	(0.4421)
0.9111	1.0000	-0.1085	-0.1595	-0.6440	-0.6942	0.6939
(0.0257)	(0.0000)	(0.0611)	(0.0507)	(0.0330)	(0.0286)	(0.4401)
0.0644	-0.1085	1.0000	0.5553	0.2021	0.0597	1.6720
(0.0595)	(0.0611)	(0.0000)	(0.0487)	(0.0640)	(0.0641)	(0.4738)
0.0289	-0.1595	0.5553	1.0000	0.0516	-0.1506	1.5004
(0.0517)	(0.0507)	(0.0487)	(0.0000)	(0.0595)	(0.0547)	(0.4718)
-0.5209	-0.6440	0.2021	0.0516	1.0000	0.2831	2.1323
(0.0509)	(0.0330)	(0.0640)	(0.0595)	(0.0000)	(0.0517)	(0.4469)
-0.6192	-0.6942	0.0597	-0.1506	0.2831	1.0000	2.6608
(0.0400)	(0.0286)	(0.0641)	(0.0547)	(0.0517)	(0.0000)	(0.4511)

Distributions a priori (i.i.d et Metropolis) et a posteriori de l'angle d'orientation θ,

pour les différents groupes de variables



Conformément aux résultats du chapitre précédent, la capacité d'estimation a posteriori du modèle est confirmée.

Il convient également de noter que, avec un nombre d'observations $T = 150$, la moyenne empirique des matrices simulées a posteriori s'approche rapidement du paramètre d"intérêt, même avec un nombre d'itérations M réduit. Pour illustrer ce fait, nous reprenons

c-dessous les moyennes empiriques de différentes simulations a posteriori, avec respectivement un nombre d'itérations M égal à 500, 1000, 2500, 5000 et 10000.

Simulation a posteriori de matrices de corrélation: $M = 500$

Espérance empirique

Matrice de corrélation						θ
1.0000	0.8954	0.1297	0.0205	-0.4760	-0.5475	1.1044
(0.0000	(0.0314)	(0.0675)	(0.0600)	(0.0490)	(0.0471)	(0.4378)
0.8954	1.0000	-0.0737	-0.2028	-0.5692	-0.6429	1.2255
(0.0314)	(0.0000)	(0.0651)	(0.0564)	(0.0420)	(0.0375)	(0.4462)
0.1297	-0.0737	1.0000	0.5300	0.2221	0.0694	1.5715
(0.0675)	(0.0651)	(0.0000)	(0.0598)	(0.0676)	(0.0688)	(0.4763)
0.0205	-0.2028	0.5300	1.0000	-0.0057	-0.1454	1.5902
(0.0600)	(0.0564)	(0.0598)	(0.0000)	(0.0648)	(0.0705)	(0.4762)
-0.4760	-0.5692	0.2221	-0.0057	1.0000	0.1924	1.5702
(0.0490)	(0.0420)	(0.0676)	(0.0648)	(0.0000)	(0.0550)	(0.4369)
-0.5475	-0.6429	0.0694	-0.1454	0.1924	1.0000	2.0503
(0.0471)	(0.0375)	(0.0688)	(0.0705)	(0.0550)	(0.0000)	(0.4415)

Simulation a posteriori de matrices de corrélation: $M = 1000$

Espérance empirique

Matrice de corrélation						θ
1.0000	0.9032	0.1234	0.0371	-0.5255	-0.6151	0.5769
(0.0000)	(0.0331)	(0.0551)	(0.0579)	(0.0451)	(0.0326)	(0.4376)
0.9032	1.0000	-0.0700	-0.1507	-0.6090	-0.7104	0.7005
(0.0331)	(0.0000)	(0.0577)	(0.0521)	(0.0506)	(0.0304)	(0.4445)
0.1234	-0.0700	1.0000	0.5583	0.2586	0.1130	1.3852
(0.0551)	(0.0577)	(0.0000)	(0.0644)	(0.0552)	(0.0634)	(0.4755)
0.0371	-0.1507	0.5583	1.0000	-0.0855	-0.0712	1.4453
(0.0579)	(0.0521)	(0.0644)	(0.0000)	(0.0687)	(0.0582)	(0.4671)
-0.5255	-0.6090	0.2586	-0.0855	1.0000	0.3574	2.1917
(0.0451)	(0.0506)	(0.0552)	(0.0687)	(0.0000)	(0.0514)	(0.4403)
-0.6151	-0.7104	0.1130	-0.0712	0.3574	1.0000	2.0087
(0.0326)	(0.0304)	(0.0634)	(0.0582)	(0.0514)	(0.0000)	(0.4536)

Simulation a posteriori de matrices de corrélation: $M = 2500$

Espérance empirique

Matrice de corrélation						θ
1.0000	0.9063	0.0932	0.0132	-0.5031	-0.5785	0.3033
(0.0000)	(0.0143)	(0.0682)	(0.0672)	(0.0494)	(0.0371)	(0.4317)
0.9063	1.0000	-0.1138	-0.1905	-0.5639	-0.6574	0.4742
(0.0143)	(0.0000)	(0.0651)	(0.0650)	(0.0429)	(0.0325)	(0.4415)
0.0932	-0.1138	1.0000	0.5517	0.2222	0.1115	1.4066
(0.0682)	(0.0651)	(0.0000)	(0.0392)	(0.0672)	(0.0645)	(0.4808)
0.0132	-0.1905	0.5517	1.0000	-0.0526	-0.1185	1.4140
(0.0672)	(0.0650)	(0.0392)	(0.0000)	(0.0681)	(0.0575)	(0.4735)
-0.5031	-0.5639	0.2222	-0.0526	1.0000	0.2099	2.1138
(0.0494)	(0.0429)	(0.0672)	(0.0681)	(0.0000)	(0.0633)	(0.4334)
-0.5785	-0.6574	0.1115	-0.1185	0.2099	1.0000	2.2861
(0.0371)	(0.0325)	(0.0645)	(0.0575)	(0.0633)	(0.0000)	(0.4439)

Simulation a posteriori de matrices de corrélation: $M = 5000$

Espérance empirique

Matrice de corrélation						θ
1.0000	0.9071	0.1593	-0.0308	-0.4663	-0.5720	0.1754
(0.0000)	(0.0233)	(0.0625)	(0.0643)	(0.0536)	(0.0435)	(0.4454)
0.9071	1.0000	-0.0341	-0.2259	-0.5555	-0.6639	0.3427
(0.0233)	(0.0000)	(0.0689)	(0.0627)	(0.0473)	(0.0395)	(0.4406)
0.1593	-0.0341	1.0000	0.5758	0.1268	0.1124	1.4574
(0.0625)	(0.0689)	(0.0000)	(0.0460)	(0.0711)	(0.0772)	(0.4740)
-0.0308	-0.2259	0.5758	1.0000	-0.0151	-0.0054	1.6620
(0.0643)	(0.0627)	(0.0460)	(0.0000)	(0.0664)	(0.0680)	(0.4744)
-0.4663	-0.5555	0.1268	-0.0151	1.0000	0.1691	2.1184
(0.0536)	(0.0473)	(0.0711)	(0.0664)	(0.0000)	(0.0620)	(0.4478)
-0.5720	-0.6639	0.1124	-0.0054	0.1691	1.0000	2.2049
(0.0435)	(0.0395)	(0.0772)	(0.0680)	(0.0620)	(0.0000)	(0.4315)

Simulation a posteriori de matrices de corrélation: $M = 10000$

Espérance empirique

Matrice de corrélation						θ
1.0000	0.9001	0.0863	0.0705	-0.4604	-0.6139	0.9713
(0.0000)	(0.0140)	(0.0724)	(0.0620)	(0.0478)	(0.0505)	(0.4452)
0.9001	1.0000	-0.0952	-0.1538	-0.5619	-0.6984	0.7281
(0.0140)	(0.0000)	(0.0694)	(0.0605)	(0.0425)	(0.0341)	(0.4396)
0.0863	-0.0952	1.0000	0.5329	0.2373	0.1897	2.0970
(0.0724)	(0.0694)	(0.0000)	(0.0452)	(0.0592)	(0.0588)	(0.4669)
0.0705	-0.1538	0.5329	1.0000	-0.0415	-0.1477	1.7893
(0.0620)	(0.0605)	(0.0452)	(0.0000)	(0.0642)	(0.0610)	(0.4741)
-0.4604	-0.5619	0.2373	-0.0415	1.0000	0.2967	2.1400
(0.0478)	(0.0425)	(0.0592)	(0.0642)	(0.0000)	(0.0602)	(0.4444)
-0.6139	-0.6984	0.1897	-0.1477	0.2967	1.0000	2.5630
(0.0505)	(0.0341)	(0.0588)	(0.0610)	(0.0602)	(0.0000)	(0.4425)

4.1.4 Limites

Bien que pouvant apporter une certaine diversité dans l'ensemble de la structure de corrélation a priori, le modèle de corrélations groupées ne dispose que d'une faible marge de manoeuvre pour tenir compte de toutes les variantes de structures possibles, tant en termes de corrélation intra-groupes (corrélation entre les variables d'un même groupe), qu'en termes de corrélation inter-groupes (corrélation entre les variables appartenant à des groupes différents)

Pour mieux illustrer notre propos, considérons par exemple le cas de 3 groupes de variables. Supposons que, a priori, nous pensons que les variables d'un même groupe

d'appartenance sont toutes fortement corrélées entre elles, alors que le 1^{er} groupe est fortement anticorrélé au $2^{ème}$ groupe (corrélation proche de -1); et que le $3^{ème}$ groupe soit décorrélé (corrélation quasi-nulle) avec les deux premiers groupes de variables.

Pour rendre compte de notre structure a priori de corrélation inter-groupes, il convient donc de donner aux 2 premiers groupes des distributions de probabilité concentrées autour des "pôles" de l'hypersphère (angles d'orientation proches de 0 et π), alors que le $3^{ème}$ groupe se concentre sur l'hyperplan orthogonal à l'axe du pôle nord (angle d'orientation d'amplitude $\frac{\pi}{2}$.)

Si notre spécification permet de bien rendre compte de la composante "inter-groupes" de la structure de corrélations, elle se révèle néanmoins désarmée pour refléter les croyances sur la composante "intra-groupes", particulièrement pour le $3^{ème}$ groupe qui est anticorrélé aux deux autres. En effet, étant donné la structure de distribution "vanilla" adoptée pour les vecteurs v_i (distribution uniforme sur l'hypersphère, conditionnellement à l'angle d'orientation), seules les variables dont l'orientation est "proche" de l'axe du pôle nord peuvent avoir une corrélation inta-groupes élevée. Plus l'orientation d'un groupe s'écarte de l'axe du pôle nord, plus la corrélation entre les variables de ce groupe est faible, la limite étant une corrélation espérée rigoureusement nulle pour les vecteurs uniformément distribués sur l'hyperplan orthogonal au pôle nord (orientation d'amplitude $\frac{\pi}{2}$).

En conséquence, le modèle à corrélations groupées est incapable de refléter l'hypothèse de forte corrélation intra-groupe, pour tous nos 3 groupes et, en même temps, de garantir que le dernier groupe s oit relativement décorrélé des deux autres.

Afin de pallier ce manquement, nous développons, dans le chapitre qui suit, l'esquisse d'un autre modèle capable de lever cet inconvenient en permettant une prise en compte simultanée des composantes inter-groupes et intra-groupes de l'ensemble de la structure de corrélations. C'est le "modèle de variables groupées", encore appelé "modèle pour une structure de corrélations hiérarchiques" dont nous planifions de poursuivre le développement complet, au-delà du présent rapoort de recherche.

4.2 Esquisse d'un modèle pour une structure de corrélations hiérarchiques

4.2.1 Motivation

Les modèles présentés dans le précédent chapitre se révélent restrictifs pour l'estimation bayésienne des matrices de corrélation. En effet, pour des raisons diverses et variées, il est naturel et courant d'observer des corrélations communes et similaires à l'intérieur de groupes de variables; et en même temps, des corrélations communes mais distinctes entre différents groupes de variables. Liechty (2004) en donne une illustration basée sur la structure de corrélation entre les rendements des actifs des secteurs énergetique et financier.

4.2.2 Formulation et Spécification

La formulation du modèle hiérarchique suit les grandes lignes de celle des cas précédents. Pour une matrice de corrélation C donnée, il s'agit toujours de définir une densité a priori sur un paramètre matriciel V, satisfaisant certaines conditions et tel que $VV^T = C$ (voir la section 3.2. pour de plus amples détails.)

Nous spécifions alors cette densité a priori suivant la méthodologie suivante, qui est une extension de celle du chapitre précédant, judicieusement adaptée pour tenir compte d'une certaine *"hiérarchie de groupes"* dans la structure de corrélation.

Nous disposons d'un nombre total n de variables, qui se répartissent en m de "*groupes de variables*." Chaque groupe i comporte un nombre n_i, $i = 1, 2, ..., m$ de variables. Ainsi, $\sum_{i=1}^{m} n_i = n$.

Comme dans le modèle précédent de corrélations groupées, des distributions indépendantes mais *non-identique* sont attribuées aux angles d'orientation θ_i de chacun des m groupes de variables distincts. Cependant, et c'est la clé du modèle de corrélations hiérarchiques, pour un groupe de variables i donné, les vecteurs v_{ij}, $j = 1, 2, ..., k$ de ce groupe sont tirés suivant une marche aléatoire de Metropolis sur une hypersphère de dimension r à partir d'un vecteur v_i servant d'axe de référence au groupe i.

Dans cette configuration, on a que:

- les angles θ_i, $i = 1, 2, ..., m$, qui donnent l'orientation de l'axe de référence de chacun des groupes par rapport à l'axe du pôle nord, déterminent le degré de corrélation inter-groupes;

- les angles θ_{ij}, $j = 1, 2, ..., k$ qui donnent l'orientation des différents vecteurs v_{ij} du groupe i par rapport à l'axe de référence du groupe, dans la marche aléatoire, définissent le degré de corrélation intra-groupes.

La densité d'un vecteur v_{ij}, $j = 1, 2, ..., k$ suit donc la densité suivante:

$$f(v_{ij}) = f(v_{ij} \mid v_i) f(v_i) = f(\theta_{ij}) \left[\frac{(2\pi)^{(r-1)/2}}{\Gamma(\frac{r-1}{2})} . \sin^{r-2} \theta_{ij} \right]^{-1} f(\theta_i) \left[\frac{(2\pi)^{(r-1)/2}}{\Gamma(\frac{r-1}{2})} . \sin^{r-2} \theta_i \right]^{-1}; \tag{4.9}$$

où $\theta_i = \cos^{-1}(v_i . \bar{v}_r) = \cos^{-1}(v_{i1})$ est l'angle entre les vecteurs v_i (axe de référence du groupe i) et $\bar{v}_r$ (pôle nord) et; $\theta_{ij} = \cos^{-1}(v_{ij} . v_i)$ est l'angle entre les vecteurs v_{ij} (variable j du groupe i) et v_i (axe de référence du groupe i)·

Chapitre5
Conclusion

Dans le présent rapport de recherche, nous élaborons une méthodologie qui permet la spécification de lois a priori sur les matrices de corrélation, dans le cadre de l'analyse statistique bayésienne.

La clé de voûte de notre modélisation est l'interprétation géométrique du concept de corrélation linéaire, en termes de produit scalaire usuel.

La méthodologie mise en place permet des simulations a priori et a posteriori de matrices de corrélation, offrant ainsi la possibilité d'affiner les appréhensions de départ de l'incertitude à l'égard de la structure de corrélation du processus étudié, grâce à l'information contenue dans les données observées.

Le principal résultat de notre travail est le "modèle d'une structure de corrélations communes", qui part de l'hypothèse que l'ensemble des variables à l'étude présentent mutuellement le même degré de corrélation. L'opérationalisation du modèle est donc fait en supposant une distribution indépendante et identique des hyperparamètres angulaires θ, pour les différentes variables. En extension à ce dernier, le "modèle de corrélations groupées", lui, autorise une certaine une hétérogénéité dans l'ensemble de la structure de corrélation. Cette variabilité est introduite par relaxation de l'hypothèse d'*identité* des distributions pour les hyperparamètres θ. Néanmoins, la portée pratique de cette extension est très restreinte car elle n'offre pas de degré de liberté permettant de tenir compte des

structures internes de groupes de variables au comportement de corrélation similaire et, en même temps, de garantir la cohérence quant aux dissemblances externes entre ces mêmes groupes de variables.

Afin de remédier à ces insuffisances, nous présentons l'esquisse d'un "modèle pour une structure hiérarchique de corrélations", dont le développement détaillé en termes de simulation a priori et a posteriori devrait se poursuivre au-delà du présent travail. L'idée-clé de ce modèle consiste à procéder à une modélisation "par groupes de variables" (hiérarchisation), ce qui offre un degré de liberté supplémentaire permettant de formuler des croyances a priori sur le degré de corrélation en tenant compte, à la fois, de l'homogénéité interne de chacun des groupes de variables, et de l'hétérogénéité entre ces différents groupes.

Des illustrations empiriques, basées sur des exercices de simulation, temoignent d'un excellent niveau d'efficacité de la méthodologie proposée.

Enfin, dans le prolongement du présent travail, quelques pistes de recherche nous semblent intéressantes à explorer. Au-delà d'élaborer une méthodologie complète de simulation pour le modèle hiérarchique, il conviendrait également d'analyser la possibilité de formuler d'autres types de densités a priori adaptées à notre méthodologie, au-delà de la forme particulière que nous avons utilisée. Aussi serait-il intéressant d'explorer des algorithmes de Metropolis-Hastings alternatifs à celui qui a été utilisé dans le présent travail.

Chapitre6
Références bibliographiques

CHIB, S., and GREENBERG E. (1995). Understanding the Metropolis-Hastings Algorithm. *American Statistician* **49**, 1-14.

DEY, D. K. and RAO C. R. (2005). *Handbook of Statistics, Volume 25: Bayesian Thinking, Modeling and Computation*, Amsterdam: North Holland.

GAMERMAN D. and LOPES H. F. (2006). *Markov chain Monte Carlo : Stochastic Simulation for Bayesian Inference* (2nd ed.), Chapman & Hall/CRC.

LEE, P. M., (2004). *Bayesian Statistics: An Introduction* (3rd ed.), London: Hodder Arnold.

LIECHTY, J. C., LIECHTY M. W. and MÜLLER P. (2004). Bayesian correlation estimation. *Biometrika* **91**, 1-14.

MCCAUSLAND, W. J. and PELLETIER D. (2005). Strategies for eliciting prior distributions on correlation matrices. *Manuscrit*.

SIGMON, K., (1993). *MATLAB Primer* (3rd ed.), Libre copie disponible sur Internet.

TANNER, M. A., (1993). *Tools for Statistical Inference* (2nd ed.), New York: Springer-Verlang.